Logique & Analyse **240** (2017), 449-463

CHOOSING BEAUTY

SIMON FRIEDERICH

ABSTRACT

Reasoning that takes into account self-locating evidence in apparently plausible ways sometimes yields the startling conclusion that rational credences are such as if agents had bizarre causal powers. The present paper introduces a novel version of the Sleeping Beauty problem—*Choosing Beauty*—for which the response to the problem advocated by David Lewis unappealingly yields this conclusion. Furthermore, it suggests as a general desideratum for approaches to problems of self-locating belief that they should not recommend credences that are as if anyone had anomalous causal powers. Adopting this desideratum, as the paper shows, yields uniformly plausible verdicts on the most-discussed problems of self-locating belief.

Keywords: Doomsday Argument, Sleeping Beauty, self-locating belief, anomalous causal powers

1. Introduction

Reasoning that takes into account self-locating evidence in apparently plausible ways often yields startling, sometimes spectacular, conclusions. One of the weirdest is that in some cases agents appear to be rationally entitled to reason as if they had "anomalous causal powers" (Bostrom (2001) p. 368). Bostrom proposes some scenarios where this happens, which are, incidentally, not among the most-discussed problems of self-locating belief in the literature: the three *Adam and Eve experiments* and the *UN*$^{++}$*-Gedanken experiment*, all introduced and discussed by Bostrom (2001).

Thispaperhasthreeaims: first, it highlights how apparent anomalous causal powers appear in the notorious Doomsday Argument (Gott (1993), Leslie (1996)); second, it proposes a novel version of the Sleeping Beauty (SB) problem (Elga (2000))—*Choosing Beauty* (CB)—for which the *halfer* view of the SB problem as advocated by Lewis (2001) yields apparent anomalous causal powers; third, it suggests that approaches to problems of self-locating belief are to be viewed in light of the desideratum that they should not recommend credences that are as if anyone had anomalous

doi: 10.2143/LEA.240.0.3254092.

causal powers, and it highlights that assigning credences in accordance with this desideratum yields uniformly plausible results.

2. How anomalous causal powers appear

Consider the following story due to Bostrom:

> Assume [...] that Adam and Eve were once the only people and that they know for certain that if they have a child they will be driven out of Eden and will have billions of descendants. [...] [T]hey have a foolproof way of generating a child, perhaps using advanced *in vitro* fertilization. Adam is tired of getting up every morning to go hunting. Together with Eve, he devises the following scheme: *They form the firm intention that unless a wounded deer limps by their cave, they will have a child.* Adam can then put his feet up and rationally expect with near certainty that a wounded deer—an easy target for his spear—will soon stroll by. (Bostrom (2001), p. 367)

Adam's and Eve's reasoning seems bizarre, but it is surprisingly difficult to determine what, if anything, is wrong with it. We can formalise it as follows: let H_1 be the hypothesis that some wounded deer will turn up, which means that Adam and Eve will refrain from having children and, accordingly, will remain the only humans ever to exist. So, the total number of observers ever to exist according to H_1 is $N_1 = 2$. Next, let H_2 be the alternative hypothesis that no wounded deer will turn up such that Adam and Eve, following their firm intention, will have children so that, at the end of the world many thousands, millions or billions of years later, a large number, say, $N_2 = 10^9$ of observers will have existed. Furthermore, let us use some principle of indifference (such as Bostrom's *self-sampling assumption* (SSA) (Bostrom (2001), p. 360)) according to which, conditional on *Hi* (with $i = 1, 2$), Adam and Eve should ascribe the probability $1/N_i$ to being any of the Ni observers ever to have lived according to H_i.[1] For example, their prior conditional credence of being the *n*-th observer ever to have lived ($n \leq N_1$) is:

$$cr(n|H_i) = \frac{1}{N_i} \cdot \tag{1}$$

[1] Indifference principles have received their fair share of criticism (see e.g. Weatherson (2005), Schwarz (forthcoming), but, for a partial vindication, also Manley (unpublished)), so the use of Eq. (1), is not an innocent step, and it may be tempting to blame the unattractive conclusion to be reached in Eq. (3) on the use of Eq. (1). However, as I argue further below, it seems unlikely that replacing Eq. (1) by some potentially more plausible alternative avoids all the unattractive features of the conclusion.

Now let us allow Adam to use his knowledge that he is the first observer ever to have lived ($n = 1$), which results in:

$$\frac{cr(H_1|n=1)}{cr(H_2|n=1)} = \frac{cr(n=1|H_1)cr(H_1)}{cr(n=1|H_2)cr(\mathrm{H}_2)} = \frac{N_2}{N_1}\cdot\frac{cr(H_1)}{cr(\mathrm{H}_2)} \qquad (2)$$

where the first line uses Bayes' theorem and the second uses Eq. (1) together with the fact that $n = 1$ is compatible with both H_1 and H_2. Assuming Bayesian conditioning and that Adam has a very small prior credence that a wounded deer will turn up, say, $cr(H1) = 10^{-7}$ (which means $cr(H2) = 1 - 10^{-7}$), we obtain for his rational *posteriors*:

$$\frac{cr(H_1|n=1)}{cr(H_2|n=1)} = \frac{10^9}{2}\cdot\frac{10^{-7}}{1-10^{-7}} \approx 50\ . \qquad (3)$$

So, if he uses his knowledge that he is the first human ever to have lived, Adam will be confident that a wounded deer will walk by.[2]

As Bostrom notes, this verdict on Adam's rational posterior credences is highly counterintuitive:

> We [...] have [...] the appearance of psychokinesis. If the example works, which it does if we assume SSA [i.e. Eq. (1)], it almost seems as if Adam is causing a wounded deer to walk by. For how else could one explain the coincidence? Adam knows that he can repeat the procedure morning after morning and that he should expect a deer to appear each time. Some mornings he may not form the relevant intention and on those mornings no deer turns up. It seems too good to be mere chance; Adam is tempted to think he has magical powers. (Bostrom (2001), p. 367)

Adopting Bostrom's term, I will say that Adam's posterior credences are as if he had "anomalous causal powers" (Bostrom (2001) p. 368).[3] According to Bostrom, strange though Adam's credences seem, they look less unacceptable if we realise that an important bit of evidence that *we* have—namely, that there will be many other observers besides Adam and Eve, including us—is simply unavailable to Adam. Bostrom concedes that the

[2] The conclusion persists qualitatively even if a prior much smaller than cr(H1) = 10^{-7} is used. Even if Adam is then no longer confident that a wounded deer will turn up, he will still seem overly optimistic given how far-fetched the possibility really is.

[3] Bostrom discusses two other Adam and Eve experiments, *Serpent's Advice* and *Eve's Card Trick*, which involve apparent anomalous causal powers in similar ways: in Serpent's Advice these powers have the flavour of "anomalous precognition" (Bostrom (2001), p. 367), and in Eve's Card Trick they appear as apparent anomalous "backward causation" (Bostrom (2001), p. 368). This last example demonstrates that apparent anomalous causal powers need not be forward directed in time.

recommendation that Adam should really have credences that conform to Eq. (3) is "deeply counterintuitive" (Bostrom (2002), p. 157), but he also points out that accepting them does not mean to ascribe *real* anomalous powers to Adam and Eve: "There is [...] no reason to ascribe anomalous causal powers to Adam. Eve and Adam would rationally think otherwise but they would simply be mistaken." (Bostrom (2001), p. 373)[4] Contrary to this remark, given the apparent implausibility of Adam's reasoning, it would seem preferable to reject Adam's conclusion outright—or, more generally, the type of reasoning on which it is based.

The apparently most straightforward way of doing so—rejecting the indifference principle Eq. (1)—is unpromising: for unless one assumes that $cr(n|H_2)$, as a function of n, is highly peaked around $n = 1, 2$ (which is necessary to have $cr(n = 1|H_1) \approx cr(n = 1|H_2)$), the effect that the ratio of the posteriors $cr(H_1|n)/cr(H_2|n)$ differs strongly from the ratio of the priors $cr(H_1)/cr(H_2)$ will persist, and this will suffice to reproduce Adam's conclusion in its qualitative features. Moreover, in a hypothetical situation where one knows H_2 to be true, i.e. that there are in total $N_2 = 10^9$ observers, but where one has not the faintest idea who among them one is, there is just no reason to assume with near certainty that one will be among the very first two ever to exist (as $cr(n = 1|H_1) \approx cr(n = 1|H_2)$ would require). To conclude, it is difficult to see how one might justify evaluating $cr(n|H_2)$ in a manner sufficiently different from Eq. (1) to avoid the conclusion reached by Adam in its qualitative features.

One can set up the notorious *Doomsday Argument* in analogy with Lazy Adam such that it recommends reasoning as if someone had anomalous causal powers. In its simplest version, the Doomsday Argument is also about two hypotheses H_1 and H_2 that differ on the total number of humans ever to exist (N_1 and N_2). Using numbers borrowed from (Bostrom (2001)), either $N_1 = 200$ billions or $N_2 = 200$ trillion humans are going to have lived. Let us assume that our empirical evidence suggests an optimistic assignment of probabilities $Pr(H_1) = 0.05$ and $Pr(H_2) = 0.95$, which we translate into priors $cr(H_1) = 0.05$ and $cr(H_2) = 0.95$. Finally, assume that you learn that you are the 60-billionth observer to exist, which, by analogous reasoning as in Lazy Adam, leads to the following ratio of posterior credences (with $n = 60$ billion $< N_1$):

[4] In his book (Bostrom (2002)), Bostrom offers an account that supposedly avoids these counterintuitive recommendations. The core idea of that account is to change to a more fine-grained reference class that includes not observers but observer stages ("observer moments"). One worry with respect to this proposal is that unless the very *structure* of Adam's reasoning in the above example is shown to be faulty, there remains the risk that similarly counter-intuitive conclusions may arise for any reference class, however well chosen.

$$\begin{aligned}\frac{cr(H_1|n)}{cr(H_2|n)} &= \frac{cr(n|H_1)cr(H_1)}{cr(n|H_2)cr(H_2)} \\ &= \frac{N_2}{N_1} \cdot \frac{Pr(H_1)}{Pr(H_2)} \cdot \\ &= 1000 \cdot \frac{0.05}{0.95} \\ &\approx 50 \end{aligned} \tag{4}$$

It appears that you should expect H_1 to be true even if the input probability $Pr(H_1)$ was substantially lower than the input probability $Pr(H_2)$. As the parallels between Eqs. (2) and (4) show, the Doomsday argument is analogous to the Lazy Adam scenario both in its conclusion and in the structure of the underlying reasoning.

To highlight the appearance of anomalous causal powers in the Doomsday Argument, assume that whether H_1 or H_2 holds depends on the success of a group of terrorists, who are trying to construct a pernicious machine which, if completed, would put an immediate end to humanity (and, so, make H_1 true). Fortunately, constructing this machine is difficult and the objective chance of the terrorists to succeed is a meager $Pr(H_1) = 0.05$. (We have some experience with the construction of machines that are of the same type but less pernicious, which allows us to assign this probability). If they succeed, $N_1 = 200$ billions of humans will have lived, if not, $N_2 = 200$ trillions. Based on the information that you are the 60-billionth human being to be born and using the same reasoning as in Eq. (4) you should have credence $cr(H_1|n) \approx 0.98$ that the terrorists will succeed in their work.

Earlier work on the Doomsday argument (e.g. Leslie (1996)) highlights that, if the argument is valid, we should take any factors that may cause humanity to go extinct much more serious than we would otherwise do. Looking at the argument through the lens of apparent anomalous causal powers, this warning translates into the recommendation that we should treat people trying to bring about the end of humanity—e.g. the terrorists—as if having *enhanced* causal powers and people trying to preserve humanity as if having *reduced* causal powers.

To illustrate how odd this conclusion is, consider some nerdy enthusiast of the type of machine that the terrorists try to construct. According to the Doomsday argument, if he cares more about contributing to the successful construction of such a machine than about humanity's future, joining the terrorists is an excellent strategy for him to achieve his aims—even if he could collaborate with more skilled collaborators when constructing the machine for neutral, perhaps even humanity-preserving, purposes. This recommendation seems extremely difficult to accept.

There have been many critics of the Doomsday Argument:[5] for example, Norton (2010) regards it as reflecting badly on the Bayesian methodology used to derive its conclusion; Eckhardt (1993), Bostrom (2001) (see fn. (4)) and Neal (2006), along different lines, hold that it is an artefact of an arbitrary and/or inappropriate choice of reference class.

Most interestingly for the further course of this paper, Dieks, building on earlier arguments by himself (Dieks (1992)) and Olum (2002), accepts the reasoning in Eq. (4) but proposes a different numerical evaluation of the expressions used there by identifyingthe *input probabilities* ($Pr(H_1) = 0.05$ and $Pr(H_2) = 0.95$ in our example) with the *posteriors* $cr(H_1|n)$ and $cr(H_2|n)$ rather than the *priors* $cr(H_1)$ and $cr(H_2)$. His central argument is that the input probabilities $Pr(H_1)$ and $Pr(H_2)$ translate into our rational credences when we are at least roughly aware of our birth rank, not in the absence of knowledge as to whether we live before the potential early end of humanity or after it. Correspondingly, when applied to Lazy Adam, Dieks' reasoning yields posteriors $cr(H_1|n = 1) = Pr(H_1) = 10^{-7}$ and $cr(H_2|n = 1) = Pr(H_2) = 1 - 10^{-7}$ according to which, as seems plausible, Adam should not expect any wounded deer to turn up.

The main reason why Dieks' strikingly simple proposal remains controversial is that it leads to *priors* $cr(H_1)$ and $cr(H_2)$ that differ from the input probabilities $Pr(H_1)$ and $Pr(H_2)$, which is an unattractive recommendation for example in cosmological theory choice (as highlighted by the *Presumptuous Philosopher* scenario due to Bostrom (Bostrom (2001), p. 124)). We will briefly look at this difficulty in Section 5 and consider how one may accept Dieks' proposal without encountering it. In the meantime, let us look for apparent anomalous causal powers in the Sleeping Beauty problem, which, as pointed out by Dieks and Bradley (2012), has a similar structure as the Doomsday Argument.

3. Sleeping Beauty and Choosing Beauty

The Sleeping Beauty problem as formulated by Elga goes as follows:

> Some researchers are going to put you to sleep. During the two days that your sleep will last, they will briefly wake you up either once or twice, depending on the toss of a fair coin (Heads: once; Tails: twice). After each waking, they will put you to [sic] back to sleep with a drug that makes you forget that

[5] There are others besides Leslie who, partly with reservations, defend its conclusion, notably Pisaturo (2009), Lewis (2010), and Bradley (2012), who argue (along different lines) that the conclusion is only apparently so implausible and only when viewed through the lens of the distorting and misleading characterization of H_1 as "doom soon".

> waking. When you are first awakened, to what degree ought you believe that the outcome of the coin toss is Heads? (Elga (2000) p. 143)

Opinions are split over the correct answer. The two candidate rational credences for Beauty ("you", in Elga's example) with respect to Heads are 1/2 and 1/3, both of which have substantial support in the literature. The simplest arguments in favor of the 1/3-view are the following (where, in accordance with convention, Beauty's first awakening is supposed to take place on Monday and the second, which occurs only if the coin falls Tails, on Tuesday): first, if the experiment is repeated many times, approximately 1/3 of the awakenings are Heads-awakenings; second, on the 1/2-view, if on Monday someone tells Beauty it is Monday, standard Bayesian conditioning tells her to shift her credence with respect to Heads from 1/2 to 2/3, i.e.[6]

$$cr^{-}(Heads)=1/2 \xrightarrow{Monday} cr^{+}(Heads) = cr^{-}(Heads|Monday) = 2/3. \quad (5)$$

This means that Beauty's rational credence with respect to *Heads* differs from its objective chance $Pr(Heads) = 1/2$, even though, knowing it is Monday, Beauty is now fully oriented about her temporal position, in apparent contradiction with David Lewis' famous Principal Principle (Lewis (1980)).[7]

The most important argument against the 1/3-view is that, in analogy with Dieks' response to the Doomsday Argument, it exemplifies a type of reasoning that yields implausible conclusions in cosmological theory choice. These will be briefly discussed in Section 5 of this paper. The essential analogy

[6] I use "cr^{-}" to denote Beauty's credences on Monday before she knows it is Monday and "cr^{+}" to denote her credences when she does know it is Monday.

[7] Elga presents his arguments in favor of the 1/3-view in (Elga (2000)). Lewis' response supporting the 1/2-answer and rejecting the argument based on the Principal Principle is given in (Lewis (2001)). The essential statement of Lewis' response is that Beauty acquires inadmissible evidence when she learns that it is Monday, which disqualifies straightforward use of the Principal Principle. This statement seems somewhat surprising, since when Beauty learns that it is Monday what essentially happens is that she ceases to be disoriented about her temporal location, hardly an uncontroversial instance of inadmissible evidence acquisition. The matter remains controversial, however, for a strong case on Lewis' behalf, see Bradley (2011).

It is impossible to do justice to the by now very extensive literature on the Sleeping Beauty problem, see (Titelbaum (2013b)) for a condensed overview. For some defences of the thirder position, see (Dorr (2002), Horgan (2004), Hitchcock (2004), Draper & Pust (2008), Titelbaum (2008), Schulz (2010), Titelbaum (2013a)), for some criticisms of it, sometimes combined with an endorsement of Lewisian halfing, see (Jenkins (2005), White (2006), Bradley & Leitgeb (2006), Bradley (2011; 2012)). Further important studies include (Kierland & Monton (2005), Briggs (2010), Ross (2010), Schwarz (forthcoming)). Articles that defend or criticise the third main position on Sleeping Beauty, the so-called "double-halfer" position, are briefly addressed in the main text further below.

between the thirder position on Sleeping Beauty and Dieks' response to the Doomsday Argument is that both identify the input probabilities—in Beauty's case the chances *Pr*(*Heads*) and *Pr*(*Tails*), in the Doomsday case the probabilities $Pr(H_1)$ and $Pr(H_2)$—with the credences one should have *when*, not before, one has the relevant bits of self-locating information "it is Monday" and "n is my birth rank". Conversely, the halfer position on Sleeping Beauty and the Doomsday Argument identify the input probabilities with the credences one should have *before*, not when, one has the self-locating information. Both Dieks and Bradley highlight these connections between viable positions on Sleeping Beauty and the Doomsday Argument (Dieks (2007) and Bradley (2012)), Dieks while defending the thirder position and rejecting the Doomsday Argument, Bradley while defending the Lewisian halfer position and endorsing the Doomsday Argument.

In addition to the thirder position as defended by Elga and Dieks and the halfer position as defended by Lewis and Bradley there is an additional, third, position on Sleeping Beauty. Its adherents Halpern (2005), Bostrom (2007), Meacham (2008), Cozic (2011) concur with Lewisian halfers that Beauty's credence with respect to Heads should be 1/2 when she awakes, but they also claim that it should *remain* 1/2 when she learns that it is Monday. I have little to say about this interesting alternative to thirding and Lewisian halfing, except that it faces the following serious problem (Bradley (2011), Titelbaum (2012)): if a coin is tossed on Tuesday evening in addition to the first one (tossed on Sunday or Monday evening), then, as is easily shown, according to the double-halfer position Beauty's credence with respect to "Today's coin will fall *Heads*" when awakening must be larger than 1/2. However, when she learns what day it is, her rational credence with respect to this proposition drops to 1/2, no matter what she learns. This makes Beauty's epistemic position oddly unstable before she is informed what day it is.[8]

Given the similarities between Lewisian ("non-double") halfing and the Doomsday Argument, can we construct a version of the Sleeping Beauty problem in which Lewisian halfing recommends credences that are as if someone (say, Beauty herself) had anomalous causal powers? We can, as the *Choosing Beauty* problem shows:

> *Choosing Beauty (CB):* As in the original Sleeping Beauty problem, Beauty is woken either once (on Monday) or twice (on Monday and Tuesday), depending on the outcome of a fair coin toss (one awakening if the coin comes up *Heads*, two if it comes up *Tails*). All her memories of any previous awakenings during the trial are erased by a drug whenever she is put to sleep. This time, however, *two* coin tosses are performed, both on Monday evening. After having been woken on Monday, Beauty is told that it is Monday and is asked to choose whether the

[8] See Conitzer (2015) for a further strong criticism of double-halfing.

> outcome of the first or the second coin toss to be performed the same evening is to count as relevant for whether or not she is woken on Tuesday. In accordance with the outcome of that coin toss, she is woken or not woken on Tuesday.

Let us refer to Beauty's two possible choices as C_1 ("The first coin toss counts") and C_2 ("The second coin toss counts"). Now consider one of the coin tosses, say, the first, and consider Beauty's rational credences with respect to its possible outcomes $Heads_1$ and $Tails_1$, as assessed from the point of view of Lewisian halfing. According to this position as applied in the original SB problem, Beauty's rational credence with respect to the outcome *Heads* on the chosen coin is 1/2 on Monday morning and 2/3 after she has been told that it is Monday. There is no reason to suppose that her rational credences about the possible outcomes of the coin toss she does *not* choose are at any stage different from 1/2. To conclude, by the standards of Lewisian halfing, after Beauty has been told that it is Monday, her rational conditional credences with respect to *Heads*1 are $cr^+(Heads1|C1) = \frac{2}{3}$ and $cr^+(Heads1|C2) = \frac{1}{2}$, and similarly (mutatis mutandis) for the other possible outcomes of the two tosses.

What seems odd about these credences is not only that there is *some* future (or past, if the coins are tossed on Sunday) coin toss with respect to which, as in the original SB puzzle, $cr^+(Heads) = 2/3 = Pr(Heads) = 1/2$, but that the identity of this coin toss (which one it is) depends on a choice Beauty makes at the very same stage. The thirder position concurs with Lewisian halfing that there is *some* stage at which Beauty's credence with respect Heads for the chosen coin should depart from 1/2 in that, according to thirdism, $cr^-(Heads_1|C_1) = 1/3$ and $cr^-(Heads_2|C_2) = 1/3$ for her credences before she is told it is Monday. However, this is not a situation where she can be given the choice between C_1 and C_2, for giving her the choice means telling her it is Monday. On Tuesday, the coin toss whose outcome decides whether she is woken once or twice has already been tossed (and, if she has been woken, fallen *Tails*). So, giving her the choice between C_1 and C_2 tells her it is Monday and, thereby, lets her credence shift to $cr^+(Heads1) = cr^-(Heads_1|C_1 \wedge Monday) = cr^-(Heads_1|C_2 \wedge Monday) = 1/2$. Accordingly, as soon as Beauty *can* make her choice, her rational credences about possible outcomes are all equal to 1/2, so that, unlike according to Lewisian halfing, there is no stage at which she simultaneously has the choice between C_1 and C_2 and a rational credence with respect to *Heads* that differs from 1/2 for some toss.

4. Biting the bullet?

As already noted, credences that are such as if someone had anomalous causal powers appear weird and counterintuitive. But perhaps this appearance

is misleading. Perhaps it is sometimes rational to have such credences even though they appear odd on superficial reflection. To pursue this suggestion, let us explore a bit further the consequences of Lewisian halfing in CB.

In order to make them most vivid, it is useful to have in mind the "extreme" version of CB ("Extreme Sleeping Beauty", Bostrom (2007), p. 66), where, if the chosen coin comes up Tails, Beauty is woken not twice, but N times on subsequent days for some very large number $N \gg 1$. In that scenario, according to Lewisian halfing, Beauty's rational conditional credences when she learns it is Monday are $cr^{+}(Heads_1|C_1) = \frac{N-1}{N} \approx 1$ and $cr^{+}(Tails_1|C_1) = \frac{1}{N} \approx 0$ and, trivially, $\mathrm{cr}^{+}(Heads_1|C_2) = cr^{+}(Tails_1|C_2) = \frac{1}{2}$ (and equivalently under exchange of the indices 1 and 2).

What makes Beauty's credences odd here is that they are analogous to those of a person who is in a position to choose between two coins to be tossed as to which of them should be *manipulated* (by affecting its internal mass distribution, say), such that the outcome of its toss becomes almost certainly *Heads*. By way of manipulating the coin, such a person would be able to *causally influence* the outcome of its toss, and the parallel between that person's rational credences and Beauty's according to Lewisian halfing confirm that the latter are indeed as if Beauty had anomalous causal powers in the sense discussed. In particular, just as it would be rational for that person to manipulate the first coin to be tossed by modifying its internal mass distribution if she wanted its outcome to be *Heads*, according to Lewisian halfing it would apparently be rational for Beauty to make the choice C1 if she wished that the outcome of the first coin toss should be *Heads*1. Accordingly, yet implausibly, putting oneself in the same situation as Beauty and choosing the first coin would be practically equivalent with manipulating it directly.

To avoid this unattractive recommendation, proponents of Lewisian halfing may appeal to causal decision theory. More specifically, they might suggest that even though Beauty's rational credences are as if she had anomalous causal powers, these odd credences are not the ones that should guide her actions and decisions. To argue for this, Lewisian halfers might compare Beauty's situation in CB to that of a subject in a medical Newcomb problem. In a typical such problem, there is some disease B for which bodily feature A is a symptom, such that A's first appearance reliably indicates that the person will fall ill with B some days later. Given the available statistical data, A and B are positively probabilistically correlated in that $Pr(B|A) > Pr(B)$, which manifests itself in the subject's rational credences $cr(B|A) > cr(B)$. This correlation, however, is not due to A's *causing* B, but, instead, due to there being some bodily state C—the presence of certain bacteria in the organism, say—that typically leads to both A and B and that *screens off* A from B in that $Pr(B|A \wedge C) = Pr(B|C)$.

Evidently, for a subject that faces a medical Newcomb problem, taking precautions against A that are not effective against C is an ineffective strategy for avoiding B. What makes medical Newcomb problems philosophically challenging is the question of whether standard (evidential) decision theory gives correct recommendations for rational action in them or whether an alternative *causal* decision theory is needed.[9] This debate aside, there is nothing particularly mysterious about them: it is unsurprising that medical Newcomb-type scenarios arise in practice, and it is uncontroversial that the rational course of action in them is not to combat the symptom A, at least not without also fighting the cause C.

I have speculated that proponents of Lewisian halfing might try to accommodate the apparently odd consequences of their position with respect to CB by appealing to causal decision theory and conceiving of CB as in essential respects analogous to a medical Newcomb problem. The crucial parallels between both cases are: first, that an agent can supposedly control some variable—the presence of the symptom A in the medical Newcomb problem and the outcome of the choice between C_1 and C_2 in CB; second, that the value of that variable is probabilistically correlated with some later event—the disease B in the medical Newcomb problem and the outcome of the coin toss chosen by Beauty in CB; and, third, that there is no causal influence from the controllable variable to the later event.

Pointing out these parallels, proponents of Lewisian halfing might argue that Beauty in the CB scenario should regard the outcome of her choice between C_1 and C_2 as merely *symptomatic* of the outcome of the chosen coin toss, just as the subject in a medical Newcomb problem should regard the symptom A as symptomatic of whether she or he will fall ill with B some days later. And indeed, this seems to be Bostrom's perspective on Lazy Adam, with respect to which he recommends that Adam may regard his "choice [as] an *indication* of a coincidence" (Bostrom (2001), p. 371), namely one between the outcome of the choice itself and the later course of events. So, given all these parallels, is the CB scenario as seen from the perspective of Lewsian halfing perhaps no more odd and problematic than a medical Newcomb problem?

Arguably not, for at some point the parallels end. In a medical Newcomb problem, correlations are non-mysterious and rational actions uncontroversial due to there being the state C which, as explained, *screens off* A from B in that $Pr(B|A \wedge C) = Pr(B|C)$. If medical research finds no state C with the required properties, the conditions for a medical Newcomb problem are not met, and taking precautions against A is (defeasibly) considered an effective means for preventing B. In the CB scenario, Lewisian halfers

[9] See (Lewis (1981)) and (Price (1986)) for examples of important contributions on the two different sides of the debate.

cannot point to any state or event C such that $cr^{+}(Heads_1|C_1 \wedge C) = cr^{+}(Heads_1|C_1 \wedge C)$, i.e. there is just no reason to expect screening off between C_1 and $Heads_1$ as far as Beauty's rational credences are concerned. So, unlike a subject in a medical Newcomb problem, if Beauty accepts the Lewisian halfer's recommendations, she has no comparable reasons to not take the probabilities $cr^{+}(Heads_1|C_1)$ and $cr^{+}(Heads_1|C_2)$ as the ones to base her rational actions on. This suggests that, according to Lewisian halfing as applied to the CB scenario, not only Beauty's rational credences but also her rational actions are as if she had anomalous causal powers.

Given these implausible consequences of Lewisian halfing when applied to CB, proponents may suggest that their position is correct only for SB but not for CB. This does not seem to be an attractive reaction, however, because there is little independent motivation to treat SB and CB differently. If Lewisian halfers choose it nevertheless, this is highly interesting and an important clarification of their position.

5. Conclusion and outlook

I conclude by offering some more general remarks on how considerations on apparent anomalous causal powers may be used to shed light on problems of self-locating belief.

The contrast between halfer-and thirder-style reasoning, to recapitulate, can be set up as a contrast between different ways of basing one's credences on input probabilities $Pr(H_1)$ and $Pr(H_2)$: halfer-style reasoning identifies the credences $cr(H_1)$ and $cr(H_2)$ themselves with $Pr(H_1)$ and $Pr(H_2)$; thirder-style reasoning, in contrast, identifies the *conditional* credences $cr(H_1|n)$ and $cr(H_2|\mathrm{n})$ with $Pr(H_1)$ and $Pr(H_2)$ (where n denotes self-locating information such as birth rank in the Doomsday Argument or day of the week in SB).

The considerations offered in the previous sections can be seen as implicitly suggesting a specific desideratum for how to form one's credences in the light of given input probabilities: the resulting credences should not be as if anyone had anomalous causal powers. As demonstrated in the previous sections, assigning credences in accordance with this desideratum yields consistently plausible results for all problems discussed: Adam and Eve cannot be confident that a wounded deer will appear; the Doomsday Argument is invalid; Sleeping Beauty should reason as recommended by the thirder response, which is excellently motivated along independent lines.

The desideratum that no one should have credences that are as if anyone had anomalous causal powers sheds an interesting light on scenarios in which thirder-style reasoning leads to unattractive conclusions. What makes it unattractive in such scenarios is its general preference for hypotheses that

predict more observers over hypotheses that predict less. (In SB, it prefers more "observers"— in the sense of awakenings—by preferring *Tails* over *Heads* in that $cr^{-}(Tails) = 2/3$ and $cr^{-}(Heads) = 1/3$.) In cosmological theory choice, for example, the general maxim to prefer cosmological theories that predict the largest possible numbers of observers does not seem plausible.[10]

The crucial point for our present purposes is that an observer in cosmology who identifies her credences $cr(H_1)$ and $cr(H_2)$ with the input probabilities $Pr(H_1)$ and $Pr(H_2)$ (by whatever means she has arrived at the latter) does not thereby violate our desideratum: unlike the credences of an observer who sets $cr(H_1) = Pr(H_1)$ and $cr(H_2) = Pr(H_2)$ in the Doomsday Argument or the CB problem, her credences are not as if anyone had anomalous causal powers. Since cosmological theories do not depend for their correctness on any agent's actions or choices, our degrees of belief in them cannot possibly be as if any agent had anomalous causal powers (which, by analogy, with the others problems discussed, would have to be powers to make some cosmological theory true). Thus, the desideratum to avoid credences that are as if anyone had anomalous causal powers does not give us any reason to adopt thirderstyle reasoning in cosmology, where it would lead to a general preference for cosmological theories that predict large numbers of observers. Encouragingly for thirders about SB, this points to a salient difference between SB and cosmological theory choice, which suggests that one can coherently be a thirder about SB without committing oneself to an unappealing principled preference for observer-rich cosmological theories.

Acknowledgements

I would like to thank various anonymous referees and the participants of my research seminar on self-locating belief at Göttingen University in the winter term 2012/13, where the ideas presented here were originally developed. Furthermore, I would like to thank audience members in Groningen and Munich for remarks and suggestions.

References

[1] Bostrom, N. (2001), The Doomsday Argument, Adam & Eve, UN++, and Quantum Joe, Synthese, 127:359-387.

[2] Bostrom, N. (2002), *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, New York: Routledge.

[10] This is highlighted in the so-called Presumptuous Philosopher thought experiment due to Bostrom (Bostrom (2002) p. 124), see (Leitgeb (2010)) for a development.

[3] Bostrom, N. (2007), Sleeping Beauty and self-location: a hybrid model, *Synthese*, 157:59-78.
[4] Bradley, D. (2011), Self-location is no problem for conditionalization, *Synthese*, 182:393-411.
[5] Bradley, D. (2012), Four problems of self-locating belief, *Philosophical Review*, 149-177.
[6] Bradley, D. and Leitgeb, H. (2006), When betting odds and credences come apart: More worries for Dutch Book arguments, *Analysis*, 66:119-127.
[7] Briggs, R. (2010), Putting a value on Beauty, In: *Oxford Studies in Epistemology, Volume 3*, T. S. Gendler and J. Hawthorne (editors), Oxford: Oxford University Press, pages 3-34.
[8] Conitzer, V. (2015), A devastating example for the Halfer Rule, *Philosophical Studies*, 172:1985-1992.
[9] Cozic, M. (2011), Imaging and Sleeping Beauty: A case for double-halfers, *International Journal of Approximate Reasoning*, 52: 137-143.
[10] Dieks, D. (1992), Doomsday—or: the dangers of statistics, *The Philosophical Quarterly*, 42:778-784.
[11] Dieks, D. (2007), Reasoning about the future: Doom and Beauty, *Synthese*, 156:427-439.
[12] Dorr, C. (2002), Sleeping Beauty: In defence of Elga, *Analysis*, 62:292-296.
[13] Draper, K. and Pust, J. (2008), Diachronic dutch books and Sleeping Beauty, *Synthese*, 164:281-287.
[14] Eckhardt, W. (1993), Probability theory and the Doomsday argument, *Mind*, 102:483-488.
[15] Elga, A. (2000), Self-locating belief and the Sleeping Beauty problem, *Analysis*, 60:143-147.
[16] Elga, A. (2004), Defeating Dr. Evil with self-locating belief and the Sleeping Beauty problem, *Philosophy and Phenomenological Research*, 69:383-396.
[17] Garriga, J. and Vilenkin, A. (2008), Prediction and explanation in the multiverse, *Physical Review D*, 77:043526.
[18] Gott, R. (1993), Implications of the Copernican principle for our future prospects, *Nature*, 363:315-319.
[19] Halpern, J. Y. (2005), Sleeping Beauty reconsidered: conditioning and reflection in asynchronous systems, In: *Oxford Studies in Epistemology, Volume 1*., T. Gendler and J. Hawthorne (editors), Oxford: Oxford University Press, pages 111-142.
[20] Hitchcock, C. R. (2004), Beauty and the bets, *Synthese* 139:405-420.
[21] Horgan, T. (2004) Sleeping Beauty awakened: New odds at the dawn of the new day, *Analysis* 64: 10-21.
[22] Jenkins, C. S. (2005), Sleeping Beauty: a wake-up call, *Philosophia Mathematica*, 13:194-201.
[23] Kierland, B. and B. Monton (2005), Minimizing inaccuracy for self-locating beliefs, *Philosophy and Phenomenological Research*, 70:384-395.
[24] Leitgeb, H. (2010), Sleeping Beauty and eternal recurrence, *Analysis*, 70: 203-205.
[25] Leslie, J. (1996), *The End of the World: The Science and Ethics of Human Extinction*, London: Routledge.

[26] Lewis, D. (1986 [1980]), A subjectivists's guide to objective chance, In: *Philosophical Papers, Vol. II*, New York: Oxford University Press, 83-132 (originally published in: *Studies in Inductive Logic and Probability, Vol. II*, ed. Richard C. Jeffrey, Berkeley: University of California Press).

[27] Lewis, D. (1981), Causal decision theory, *Australasian Journal of Philosophy*, 59:530. Lewis, D. (2001), Sleeping Beauty: reply to Elga, *Analysis*, 61: 171-176. Lewis, P. J. (2010), A note on the Doomsday argument, *Analysis*, 70:27-30. Manley, D. (unpublished), On being a random sample, retrieved from *http://www-personal.umich.edu/~manley/Site/Home.html*, 3 July 2015.

[28] Meacham, C. (2008), Sleeping Beauty and the dynamics of *De se* belief, *Philosophical Studies*, 138:245-269.

[29] Neal, R. M. (2006), Puzzles of anthropic reasoning resolved using full non-indexical conditioning, http://arxiv.org/abs/math/0608592.

[30] Norton, J. (2010), Cosmic confusion: not supporting versus supporting not-, *Philosophy of Science*, 77:501-23.

[31] Olum, K. (2002), The Doomsday Argument and the number of possible observers, *The Philosophical Quarterly*, 52:164-184.

[32] Pisaturo, R. (2009), Past longevity as evidence for the future, *Philosophy of Science*, 76:73-100.

[33] Price, H. (1986), Against causal decision theory, *Synthese*, 67:195-212.

[34] Ross, J. (2010), Sleeping Beauty, countable additivity, and rational dilemmas, *Philosophical Review*, 119:411-447.

[35] Schulz, M. (2010), The dynamics of indexical belief, *Erkenntnis*, 72: 337-351.

[36] Schwarz, W. (2015), Belief update across fission, *British Journal for the Philosophy of Science*, 66:659-682.

[37] Titelbaum, M. G. (2008), The relevance of self-locating beliefs, *Philosophical Review*, 117:555-606.

[38] Titelbaum, M. (2012), An embarrassment for double-halfers, *Thought*, 1:146-151.

[39] Titelbaum, M. (2013), *Quitting Certainties: A Bayesian Framework Modelling Degrees of Belief*, Oxford: Oxford University Press.

[40] Titelbaum, M. (2013), Ten reasons to care about the Sleeping Beauty, *Philosophy Compass*, 8:1003-1017.

[41] Titelbaum, M. (forthcoming), Self-locating credences, In: *The Oxford Handbook of Probability and Philosophy*, A. Hájek and C. Hitchcock (eds.), Oxford: Oxford University Press.

[42] Weatherson, B. (2005), Should we respond to evil with indifference, *Philosophy and Phenomenological Research*, 70:613-635.

[43] White, R. (2002), The generalized Sleeping Beauty problem: a challenge for thirders, *Analysis*, 66:114-119.

Simon FRIEDERICH
University of Groningen,
University College & Faculty of Philosophy
The Netherlands
s.m.friederich@rug.nl
www.simonfriederich.eu